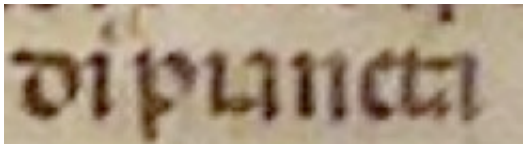
VII. Appendices: C. The *Commedia* Project Encoding System

Barbara Bordalejo

The first part of this article describes the development of the encoding system originally devised for the *Commedia* Project,[1] and subsequently adapted for other projects.[2] The second part of the article describes the encodings used for the transcription and editorial phenomena described elsewhere in this publication.

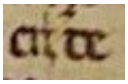
The Development of the *Commedia* Encoding System

The transcription and encoding system used in this edition are the result of a concerted effort that started in 2001.[3] Most of the decisions regarding manuscript transcription were taken early on in the project.[4] The encoding system, however, matured over a period of several years between 2001 and 2004. Over this time, careful rethinking of the aims of transcription, and of how transcripts might most usefully be encoded, led to the project adopting conventions, described here, differing markedly in certain respects from other manuscript transcription systems. Accordingly, the project's work may be of interest to other scholars engaged in manuscript transcription. I offer examples in order to facilitate understanding and further use of the system.[5]

From its beginnings, it was agreed that the *Commedia* Project's transcription protocols should be based on those of the Società Dantesca for their Dante Online website (http://www.danteonline.it/english/risorse.htm). Indeed the structure of the internal document which was used as a basis for the *Commedia* transcriptions follows the original order of elements as laid out in the Società Dantesca's website. These guidelines take into consideration practical matters concerning spellings, punctuation, word division and the expansion of abbreviations, and they also offer a form of symbolic representation – based on conventions – to convey the transcriber's interpretation of what he or she believes to be in the manuscripts. For example, the Società Dantesca transcribes a correction in ms. Riccardiana 1005, *Inf.* i 17 in this way[6]:  <di +i0 del>

These symbols are used to represent a correction. In this case, the correction was carried out by the main scribe of the text – or by a hand which can not be distinguished from the main hand – indicated by 0. The complete set of symbols is enclosed in angle brackets. The first word, in this case "di" is the one which was originally in the manuscript, and the last word – "del" – is the one which replaced it. Next to the 0 (representing the main hand or one which cannot be distinguished from it) the plus symbol is used, denoting addition, followed by the letter "i" which indicates that the correction has been introduced

between the lines, i.e. it is interlinear. The Società Dantesca guidelines allow the possibility of marginal additions – "m" – or additions within the line – for which they do not use any symbol. In this specific case, according to the transcription produced by the Società Dantesca, the manuscript has the word "di" which has been substituted by the word "del," creating the phrase "del pianeta" instead of the original reading "di pianeta."

A second example can be found in Riccardiano 1005, *Inf.* i 94:  <crede +i0 cride>

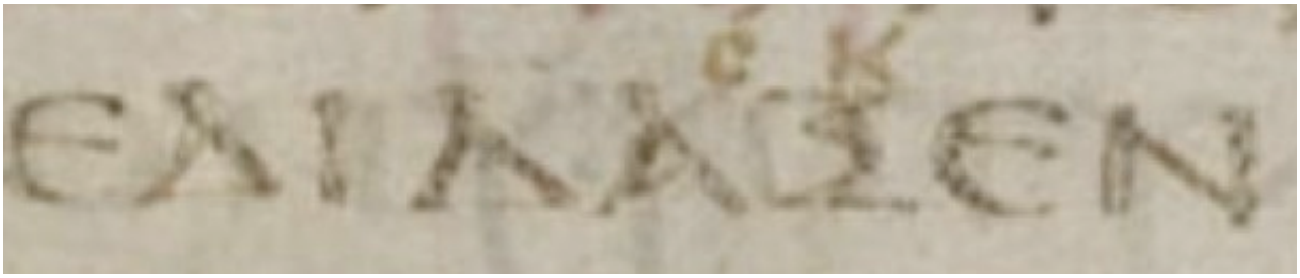
Here, the original reading "crede" is followed by the identifiers for the position and the scribe, and at the end, the modified reading "cride," again, all enclosed in angle brackets.

Although the Società Dantesca's guidelines were useful as a basis for the *Commedia* Project's transcription protocols, a new encoding system was required in order to record not only that changes had been introduced to the manuscripts, but also more specific details about how these changes came to be. As they stand, the Società Dantesca's guidelines provide information about what the editor believes to be the original reading and the final reading in a document. Consider again the above example <crede +i0 cride>. This describes the first reading in the document (crede) and the later reading (cride). It also tells us that the second reading is the result of an interlinear (i) insertion (+) and that the correction was carried out by the same scribe or by a hand that cannot be distinguished from his. However, the information that we are given about how the correction was carried out is incomplete. From the encoding alone it is not possible to know whether the whole of the word "cride" was written between the lines or whether it was only the letter "i." The system also offers no information about how or if the original reading was cancelled.

At the same time as we were considering these issues, two major Greek New Testament editing projects were exploring the same problem, of multiple levels of correction within a particular witness.[7] There were close informal connections between these projects and the *Commedia* project, which led to the development of a common approach to the problem. For the New Testament editorial communities, the preservation of the record of corrections within a single witness is critical evidence. Multiple readings in a witness might be evidence of contamination from a different manuscript group, a well-known phenomenon in New Testament editing particularly among later manuscripts. In other cases, alternative readings within manuscripts might be the only extant traces of otherwise lost texts.

Accordingly, the New Testament projects had been routinely recording alternative readings within manuscripts. By early 2001, at the time we commenced work on the *Commedia* encoding, the New Testament projects had implemented a formal encoding for variants within a document, by the use of the TEI <app> (for "apparatus") element in the encoding of individual witnesses.[8] This is a standard TEI-

XML element used to separate multiple variant readings occurring in a particular place of variation in an individual witness. For example, when a manuscript was amended by its main scribe or by a later corrector, both readings are included as part of the transcription. As employed by the IGNTP and by the Institute for New Testament Studies, there are at least two reading elements in each particular example. In an example from Codex Sinaiticus, quire 66, Folio 5r, first column, line 5 (http://www.codexsinaiticus.org), we find that the reading εδιδαξεν has been corrected to εδιδαcκεν:



This correction is expressed in XML as follows:

```
<app>
<rdg type="main-corr"><w n="11">εδιδαξεν</w></rdg>
<rdg type="corr" n="ca"><w n="11">εδιδαcκεν</w></rdg>
</app>
```

The transcription makes no attempt to represent the document and it does not include the standard <add> and <del> elements.[9] The recorded readings, both included within the <app> element, are explicit declarations of different states of the text as perceived by the editors and are presented as complete and meaningful entries. These readings are particularly useful in the context of collation and for the production of a critical apparatus. This approach prioritizes editorial opinion and takes no notice of the documentary aspects of the text.

The encoding system developed for New Testament projects was a useful place to start when I was devising the *Commedia* Project's transcription and encoding guidelines. Its main drawback was that while these projects used the <rdg> elements within <app> to give the variant states of the text, they gave no information about the text of the document.

In the preceding sentences, I have introduced a distinction between "the text of the document" and the "variant states of the text". Because this distinction is so crucial to what follows, and may be unfamiliar to the reader, it needs further explanation. In this article, I use the phrase the "text of the document" to refer to the sequence of marks present in the document, independently of whether these represent a complete, meaningful text. That is: the reader sees a sequence of letters, occurring in various places in relation to each other (perhaps between the lines or within the margins) and carrying various markings (perhaps

underdottings or strikethroughs). These make up what I here refer to as the text of the document.

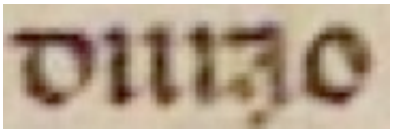
The reader understands the marks present in the text of the document as meaningful and constructs one or more specific senses from them. Where more than one sense can be constructed from the text of the document, I refer to these as the "variant states of the text", or as the "constructed" texts. I deliberately avoid the use of the phrase "the text of the work," as this is a completely different concept that refers specifically to an abstract concept of "the work." (Cf. Tanselle, *The Rationale of Textual Criticism*.) In our system, at each point of variation the text of the document at that point is encoded in a <rdg type="lit"> element. The variant texts, the constructed texts, are encoded within <rdg type="orig"> <rdg type="c1"> <rdg type="c2"> elements.

Over the course of many conversations with the New Testament scholars, we learnt that for them, the variant states of the text were of crucial importance. However, how those variant states were actually represented in the document – the text of the document – was of much less interest. Therefore, at places of variation in the manuscript they commonly encoded the variant states of the text, but said nothing about how the text of the document actually appeared at that point. Thus, in the Sinaiticus example above the alternative readings εδιδαξεν and εδιδαCκεν are recorded, giving the variant states of the text at this point. However, there is no attempt to record exactly how the text appears on the page (with two letters Cκ written in a different hand above the ξ): that is, the text of the document.

From the first, the *Commedia* Project determined that it was crucial to record the text of the document as well as the variant states of the text. While the scholarly community tends to accept editorial opinion as fact, it does not follow that editors' interpretations are always correct. Hence, this project (and others in which I was involved) felt it important to record the text of the document as well as the variant states of the document. To do this, we introduced an additional <rdg> element with a different attribute, one that would attempt to make explicit the exact sequence of meaningful letters and markings in the document. This was another <rdg> element, but with the type attribute set to "lit": <rdg type="lit">. Unlike the other <rdg> elements within <app>, <rdg type="lit"> would contain the closest representation of the "text of the document": the sequence of meaningful marks on the original document. Its objective was to aid the reader in the interpretation of a manuscript and to allow the possibility of a different interpretation from that of the editor.

In its first incarnation, what came to be affectionately called by the *Commedia* transcribers the "literal tag" or "lit tag" included the standard TEI elements <add> and <del>. However, it was decided that these elements are inappropriate within a representation of the text of the document.[10] The <add> and <del> elements combine a statement about the variant states of the text (that is, about the text before and after the change)

and about the text of the document (that is, about the letters and marks present on the page). “Addition” and “deletion” are not something that happen in a document, but are better described as the human interpretation of the text of the document, based on the reader’s understanding of the methods used by authors and transcribers to modify text.[11] Certain acts in any writing process are understood by readers as deletions. A crossed out text is understood as deleted and so is an erased one or a scraped one. Underdotting can be understood as deletion, or the text might have a tiny “vacat” written around it. Occasionally a particular word is understood as deleted because it is clear that it is meant to be replaced by a different one, even when there are no signs to mark this deletion at all.[12] These acts are all interpretive, as the predictable behaviour of someone (an editor, a transcriber or a reader) who frequently encounters those signs.

Consider the following example from ms. Riccardiana 1005, *Inf.* iii 9: 

Our first attempt to encode this, before we came to see that the <add> and <del> elements were inappropriate when representing the text of the document, was:

```
<app>
<rdg type="orig">dura</rdg>
<rdg type="c1">duro</rdg>
<rdg type="lit"><del rend="underdot">dura</del><add>duro</add><rdg>
</app>
```

The dot under the letter “a” marks a place in which a purposeful alteration has been introduced. Here, our original XML-TEI expression of this uses both <add> and <del>. What we see in the image of the manuscript is a word “durao” in which the letter “a” has been underdotted. However, when we first translated this into the newly developed system we realized that to say that the word “dura” has been deleted would not be correct, and even less correct is to say that “duro” has been added.

What happens on the page is not that the whole word “dura” has been deleted, and the whole word “duro” added: only one letter is changed, in fact. So, we considered an alternative encoding, which would show that the changes affected only one letter, thus:

```
<rdg type="lit">dur<del rend="underdot">a</del><add>o</add><rdg>
```

This appears more specific, and hence more satisfactory: only the letter “a” is underdotted and only the “o” added. But it is misleading to use the terms deletion and addition here. Firstly, the “o” is not added at all. Its appearance following the “a” is just the continuation of the normal writing process. It is as much a

distortion to say that “o” is here added as to say that when writing “the”, one first writes “t”, then adds “h” and “e”. Second: the statement that the underdotting of “a” is a deletion is not a statement about what actually appears in the manuscript. The “a” is actually not touched at all: simply, a dot is placed under it. The interpretation, that this is a deletion, is a statement about the variant states of the text, not about the text of the document.

Those familiar with medieval manuscripts, scribes and their writing practices immediately recognize the dot under the “a” as an expunction mark. Thus they read this text as follows: the scribe wrote “dura,” realized that this was a mistake and corrected the reading to “duro.” This train of thought is so ingrained that readers do not perceive it as a series of separate states. It takes only a fraction of a second to think and realize what has happened. But what occurs are indeed two distinct activities. Firstly, the reader realizes that there is a set of marks on the page that are text. Secondly, the reader constructs meaning out of those marks on the page. The first is an act combining perception and interpretation, the second is an act purely of interpretation. It was this reading that brought into question the idea of using the <add> and <del> elements. Both <add> and <del> confound the representation of the text of the document with the representation of the variant states of the text constructed by the reader.

Up to this moment of realization, the encoding of projects similar to the *Commedia* Project, such as the Canterbury Tales Project, attempted to present simultaneously both “what is in the document” as a series of additions or deletions, and “what is in the text”, as a series of distinct readings. This arose from the misunderstanding of <add> and <del> as elements that could objectively describe the text of the document. But this distinction, between the text of the document and the text that is constructed by the reader/editor, only became evident when we first tried to implement <rdg type="lit">. What seems obvious now (the distinction between the text of the document and the text or variant texts as the editor perceives it to be) required months of discussion with Klaus Wachtel (from the Institute for New Testament Research in Münster) about the transcription of corrections of the manuscripts of the Greek New Testament, before new ideas about how to encode these different reading stages started to emerge. These discussions were the base of the encoding system developed for the *Commedia* Project, used in this DVD-ROM and now implemented in other projects.

The system I devised includes a new set of parameters for the elements that should be allowed within the <rdg> element with attribute type="lit.” Only the visible, physical features of the text of the document are represented here. In the case of the example from the Riccardiana manuscript, discussed above, the resulting encoding is:

```
<app>
<rdg type="orig">dura</rdg>
```

<rdg type="c1">duro</rdg>
<rdg type="lit"> dur<hi rend="ud">a</hi>o</rdg>
</app>

Encoded in this manner, the editorial judgement, in the form of the editor's construction of the variant states of the text, is clearly articulated in <rdg type="orig"> and <rdg type="c1">, while in <rdg type="lit">, we find a more neutral expression of the text of the document. Notice that in the above example, only the letter "a" requires further encoding in the form of <hi rend="ud"> to indicate the expunction mark. The "o," which is the result of the scribe continuing to write as normal, requires no special encoding and neither do the first three letters in the word ("d," "u," "r"), which are not affected by the change.

The main goal of this new transcription system is to present a clear distinction between the text of the document (i.e. what goes in the lit tag: the exact series of marks upon the page) and how the editor (or the transcriber) interprets the different stages of development of the text (i.e. our understanding of the text as originally written and then altered). These two levels must always be clearly distinguished. Although both of them are interpretive, they are interpretive in different ways and they serve different purposes. The first attempts to show the letters and marks which appear to be present in a particular document and the second offers an opinion which explains what the editor thinks is the text or texts which can be constructed out of those marks. Both are "texts": but they are different kinds of texts. The text of the document is the sequence of letters and meaningful marks the reader sees on the page. From this, the reader constructs one or more texts. Usually, where letters follow one another into words in an uninterrupted sequence, the text of the document and the constructed text appear identical. But in cases such as this "dura/o" example, a distinct act of interpretation is required to construct the variant texts from the text of the document. Our division between recording the text of the document (in <rdg type="lit">) and recording the variant texts (in the other <rdg> elements grouped within the <app> element) makes this distinction explicit.

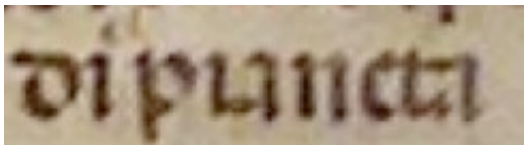
Thus, the *Commedia* Project encoding system aims both to represent the different stages of variation in the text and to give a concrete form to its expression. When a transcriber finds a "place of variation" in the manuscript, he or she can use the <app> element. This contains two main components:

(a) a sequence of two or more <rdg> elements giving the editor's interpretation of the variant states of the text, typically with <rdg type="orig">, the original reading[13] followed by <rdg type="c1">.[14] If there are more than two stages in a correction, for example, in the case of having more than one corrector, these stages are presented in what is deemed to be their successive order;

(b) <rdg type="lit">, what "literally" is in the witness; that is the text of the document.

Although the intricacy of this system can be perceived as a disadvantage, the sophistication of the final results is well worth the effort. In the final display, we can present the different scribal hand, or stages of correction by the same scribe at different points in time. This is particularly important because one of the witnesses included in this DVD-ROM is Luca Martini's copy of the Aldine edition of the *Commedia*. Martini corrected his copy against a manuscript that has since been lost. Martini's corrections become accessible thanks to the separation of the original text from the corrected version as expressed with the different <rdg type="orig"> and <rdg type="c2"> elements. The system also opens an important avenue for recording the creative process of an author, as is done in genetic editions.

I will now consider some practical cases in which this encoding system is particularly effective. For example, the Società Dantesca example offered above is rendered as follows in the *Commedia* Project:

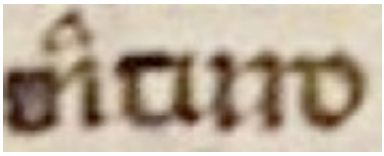


```
<app>
<rdg type="orig">di</rdg>
<rdg type="c1">del</rdg>
<rdg type="lit">di<s type="il">el</s></rdg>
</app>
```

In this example, the original reading is not altered at all; instead, the letters "el" are written between the lines in a smaller size and in what today seems a fainter ink. The transcription offers a history of what has happened here: that the main reading (the original reading in this witness) was, at some point, deemed to be incorrect and a correction in the form of an interlinear addition has been supplied. The encoding here distinguishes clearly two kinds of editorial activity. First, within the first two <rdg> elements we interpret the different texts which can be extracted from the manuscript at this point: thus "di", seen as the "original" reading, and "del", seen as the "c1" reading. It is an editorial decision to assume that the scribe meant "del" to replace "di."[15] This is expressed in the first two <rdg> elements, with the "type" attribute used to declare the agent responsible for these readings in the manuscript.

Second, within the <rdg type="lit"> element, we show what we see as the text of the document. Here, the "i" in "di" has not been assigned any specific encoding by the transcriber because it was not deleted by the scribe. Literally the manuscript reads "diel", with the "el" written above the "di". The <rdg type="lit"> element attempts to present what, seemingly, the manuscript shows: that the word "di" was written and that, at a somewhat later stage, the letters "el" were added. One could take this further and perhaps offer a theory about whether the corrections came from a manuscript representative of a different part of the

textual tradition.

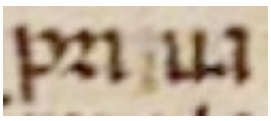
A type of correction commonly found among the witnesses of the *Commedia* transcribed and encoded for this project is the rewriting of a letter. The system considers the rewriting of one or more characters as a particular kind of replacement. In this edition, all replacements are encoded using the <s> element and placed within the <rdg> element. For example: 

```
<app>
<rdg type="orig">sicuro</rdg>
<rdg type="c1">maturo</rdg>
<rdg type="lit"><s type="rp"><s type="cow">si</s>m</s><s type="il">a</s><s type="rp"><s
type="cow">c</s>t</s>uro</rdg>
</app>
```

As the first two <rdg> elements show, the editor believes the scribe originally wrote “sicuro” and changed this to “maturo.” Here we have two examples of the scribe rewriting the original character. The first character of the word sicuro, the letter “s”, is overwritten to appear as the first two minims of the letter “m” (notice that the minim that represents the letter “i” does not suffer any alteration and yet its meaning changes because of its new context). The letter “a” appears above the word and it is reasonable to think that it was added during the revision of the text. A second instance of a replacement by overwriting is the letter “c” which has been transformed into a “t.”

The reason why the attribute of the original <s> element is type="cow" has to do with the ambiguity of the word “rewritten.” Does the word “rewritten” refer to what has been overwritten (as in “the letter c is overwritten by a t”), or to the overwriting (as in “the letter t overwrites the c”)? I decided to embrace the Saussurean idea of the arbitrariness of the sign, by using a completely random word to refer to the first state of the characters in a particular witness.[16]

The reasons for separating a letter into its smallest parts become clearer the more one looks at the scribal treatment of these. Consider this example:



The project encoded this as:

```
<app>
<rdg type="orig">prima</rdg>
```

<rdg type="c1">priua</rdg>
<rdg type="lit">pri<s type="rp"><s type="cow">m</s>u</s>a</rdg>
</app>

The problem with this encoding is that it suggests that the letter “m” was overwritten and replaced by the letter “u. ” We can all agree that the first reading in this document was “prima” and that it now reads “priua.” However, it is much more difficult to agree that this is a literal description of the manuscript.[17]

We have found that this encoding system presents several advantages. Firstly, the transcribers can defer interpretation of the stages of meaning, since the element <rdg type="lit"> can be transcribed independently of <rdg type="orig"> and <rdg type="c1">. This also allows the editor of a publication to make a final decision as to what happened at each individual place of variation. Secondly, the use of <rdg type="lit"> allows us to present a closer reconstruction of what actually appears in a document on the computer screen. Thirdly, the other components of the element <app> (<rdg type="orig">, <rdg type="c1">, <rdg type="c2">, <rdg type="c3">, etc.) can be collated separately from the rest of the text. The separate collation of multiple readings in a witness can be most useful when a scribe used a witness of different affiliation to correct his copy. In such cases, separate collation allows the isolation of readings which originated in different manuscripts and which could hint at distinct affiliations in a single text. Separate collation might also be of help in cases in which conflation has occurred because a manuscript is corrected with readings from another one from a different branch of the textual tradition.

Currently, a version of the encoding system of the *Commedia* Project has also been implemented for use by the Canterbury Tales Project and by the Cancioneros Project. It has not yet been used in textual traditions where authorial variation is present, and the advantages of this system when applied to authorial manuscripts are yet to be fully explored and exploited.[18] It should work as efficiently to distinguish different authorial versions of a particular text, which in turn should translate into an easier reconstruction of these versions and allow the distinction and separate reconstruction of different authorial stages of composition, thus permitting the creation of genetic editions.

Barbara Bordalejo

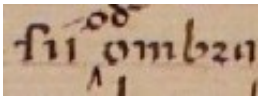
4th April 2010

The Encodings

This section gives both the Collate-style encoding used by the transcribers, and the XML encoding into which this was translated, and which is used in this publication.

Position

Interlinear: Collate [i]od[/i]; XML <s type="il"></s>

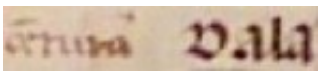


sii \od/ ombra *Inf.* i 66 LauSC

Right margin: Collate [rm][/rm]; XML <s type="rm"></s>

temesse ←\tremesse/ *Inf.* i 48 LauSC

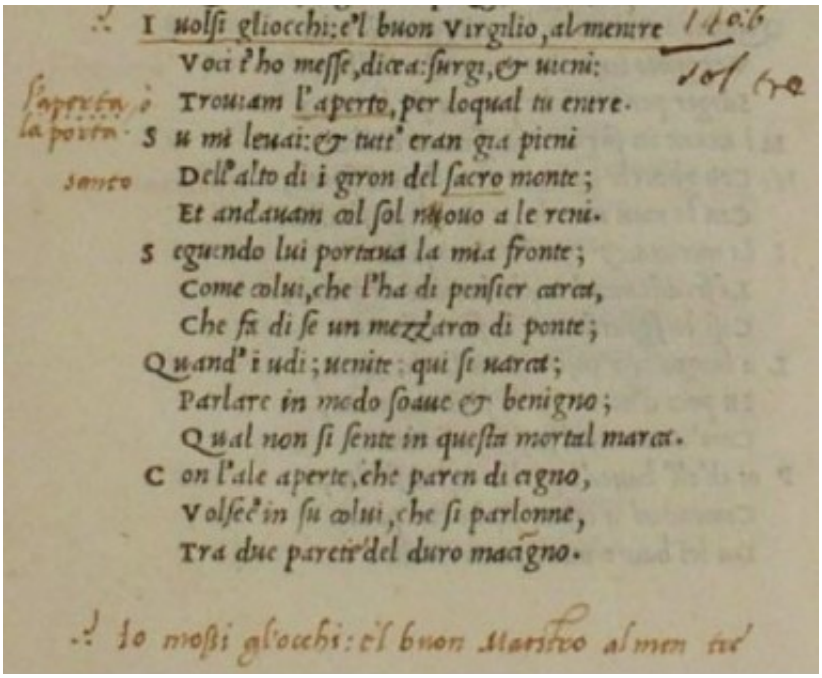
Left margin: Collate [lm][/lm]; XML <s type="lm"></s>



Da la →\centura/ *Inf.* x 33 Rb

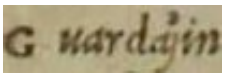
Top margin: Collate [tm][/tm]; XML <s type="tm"></s>

Bottom margin: Collate [bm][/bm]; XML <s type="tm"></s



I uolsi ↑\Io mossi/ gli occhi ↑\gl occhi/ e l ↑\e l/
buon Virgilio ↑\buon Maestro/ al mentre ↑
\al men tre/ N *Purg.* xix 34 Mart

A letter or word added within the line by cramming between words or at either end of the line but attached to it: Collate [pl][/pl]; XML <hi rend="cr"></hi>



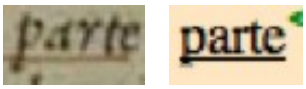
Guarda\i/ *Inf.* i 16 Mart

Scribal Deletion

Underdotted, or erased by dots within the letter or dots enclosing the word: Collate [ud][/ud]; XML <hi rend="ud"></hi>


quale *Inf.* i 55 LauSC

Underlined: Collate [ul][/ul]; XML <hi rend="ul"></hi>


parte *Inf.* vii 82 Mart

Cancelled by a stroke through the letter or the word: Collate [st][/st]; XML <hi rend="strike"></hi>


que *Inf.* v 78 Mar

Erased: Collate [er][/er]; XML <hi rend="er"></hi>

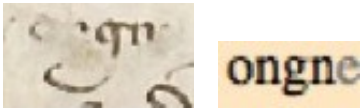

di una *Inf.* i 49 LauSC

Problematic Readings

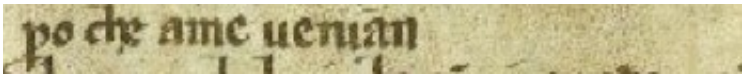
Unreadable, including words or letters missing because of physical damage to the manuscript: Collate [unr]xxx[/unr] the number of x's corresponds to the number of letters that could have been present); XML <gap extent="2"/> (the value of the extent attribute corresponds to the number of letters that could have been present)


la oue [..]icesti *Inf.* i 133 Rb

Doubtful or uncertain readings: Collate [dub][/dub]; XML <unclear></unclear>


ongne *Inf.* iii 9 Triv

Space left deliberately by the copyist, either because he is unsure of the reading or because there is a blank space in his exemplar: Collate [sp]xxx[/sp] (the number of x's corresponds to the number of letters the space could accommodate); XML <space dim="h" extent="3"/> (the value of the extent attribute corresponds to the number of letters the space could accommodate)


pero che a me uenian [ ] *Par.* xiv 125 Ham

Elements can be used together, thus for an erased reading which is unreadable: Collate [er][unr]xxx[/unr][/er] (the number of x's corresponds to the number of indecipherable letters); XML <hi rend="er"><gap

extent="3"/></hi> (the value of the extent attribute corresponds to the number of indecipherable letters)

 so[.] *Inf.* ii 3 Triv

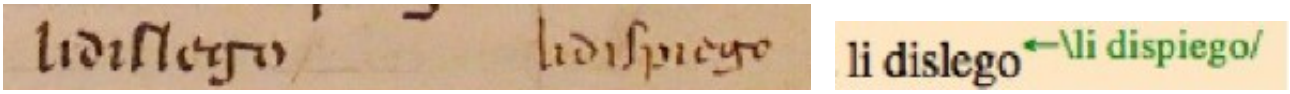
Glosses and Alternative readings

During transcription, these were encoded as notes within Collate: thus the gloss “Luxuria”in the right margin of Ham at *Inf.* i 32 was recorded as “{line 32: gloss: Luxuria}”. Later, these were converted into XML <note> elements, kept apart from the transcripts, thus:

<note id="Gl-Note-IN-1-32-Ham" type="gloss" rend="rm">Luxuria</note>

The value of the “id” attribute connects this to line 32 of Canto 1 of *Inferno* in Ham; the value of the “rend” attribute places the gloss in the right margin. The values “lm” “tm” “bm” for the “rend” attribute place the gloss in the left, top and bottom margins.

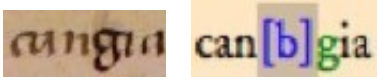
Alternative readings, as opposed to glosses, are encoded as part of the running text in the transcripts, together with information as to the location of the alternative reading: Collate [al][rm][/rm][/al] (for an alternative reading in the right margin); XML <s type="al"><s type="rm"></s></s>



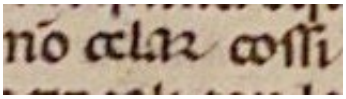
Purg. xxv 31 LauSC

Substitution of one Reading for Another

Replacement when the original reading is still visible and legible: Collate [rp][cow]abc[/cow]def[/rp]; XML <s type="rp"><s type="cow">abc</s>def<> (*abc* is the original reading, *def* is the reading which takes its place)

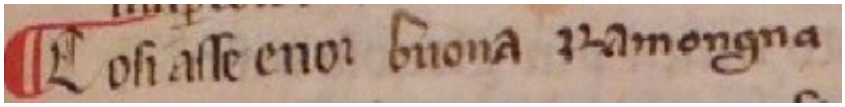
eg. LauSC *Inf.* ii 38 

Replacement over an erasure where the original reading cannot be deciphered: Collate [rp][er][unr]xxx[/unr][/er]abc[/rp] (the number of x’s indicating the number of illegible letters, abc is the reading which replaces it); XML <s type="rp"><hi rend="er"><gap extent="3"/></hi>abc</s> (the extent attribute indicates the number of illegible letters; abc is the reading which replaces it)

eg. Rb *Purg.* xxxii 5 

A word or phrase added in a space left by the copyist: Collate [rp][sp]xxxx[/sp]abc[/rp] (the number of x’s indicates the number of letters the space could accommodate, abc is the added word or phrase; XML <hi

rend="inspace">abc</hi> (abc is the word or phrase written in the space).

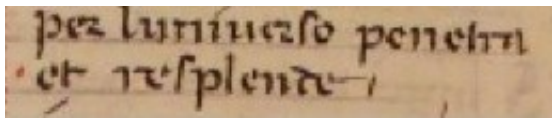
eg. Ash *Purg.* xi 25 

Cosi a sse e noi [\ buona Ramongna /]

Aspects of Layout

Superscript: Collate [sup][/sup]; XML <hi rend="sup"></hi>

Line break: Collate and XML &lb;

eg. 

per l uniuerso penetra
et resplende

Par. i 2 LauSC

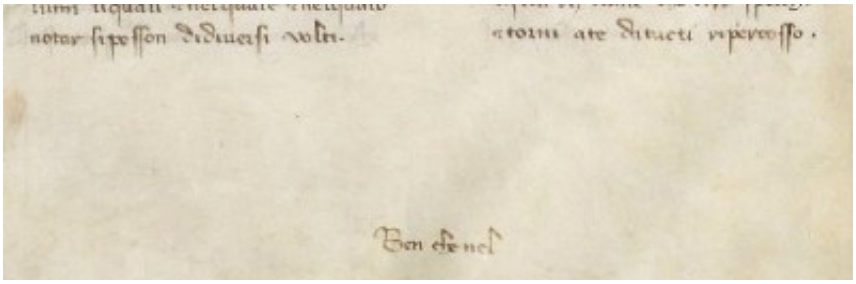
Line break with concatenation marker to indicate that a word is split across the line break: Collate and XML &lb;=

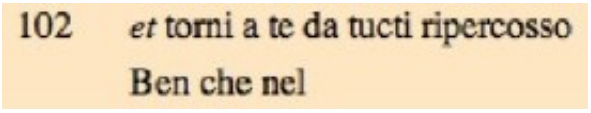
eg. 

PER COR
RER MI
GLOR
ACQUA
ALÇA LE
VELE

Purg. i 1 Triv

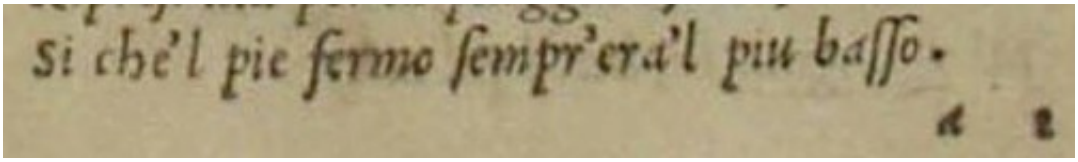
Catchword: Collate {/cw/ } (that is: within a Collate “note” structure, typed as a “cw”); XML <note type="cw"></note>





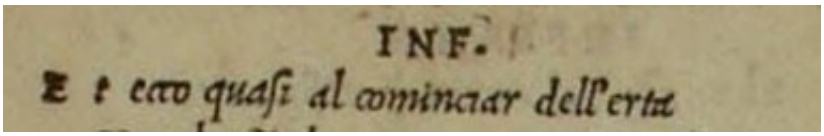
Triv *Par.* ii 102/103

A signature: Collate {/sg/ } (that is: within a Collate “note” structure, typed as a “sg”); XML <note type="sg"></note>



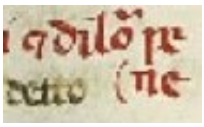
Mart *Inf.* i 30

A running head: Collate {/rh/ } (that is: within a Collate “note” structure, typed as a “rh”); XML <note type="rh"></note>



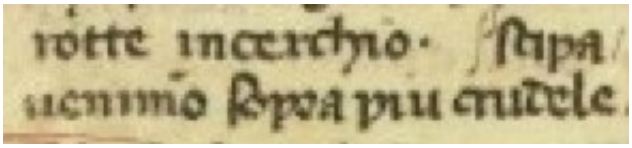
Mart *Inf.* i 31

Wrapped line below: Collate and XML &wlb;



et di lor pe↑ne *Inf.* xviii IR Ham

Wrapped line above: Collate and XML &wla;

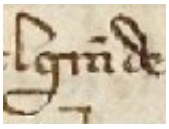


3 uenimmo sopra piu crudele ↓stipa

Inf. xi 3 Ham

Expansion of an abbreviated form, used more frequently than any other tag in the *Commedia* project: Collate [exp][/exp]; XML <expan></expan> .

et *Inf.* i 82 Triv



l grande *Inf.* i 83 Triv

Notes

1. Although the encoding described here was devised by myself and Peter Robinson, the whole

Commedia Project team contributed by producing examples and bringing to my attention new cases that had not yet been considered or instances in which, for one reason or another, the original encoding did not work. I would like to thank Peter Robinson and Prue Shaw for their suggestions about this article and the *Commedia* Project team for contributing to the development of the guidelines and for their efforts in using them to transcribe the witnesses of the *Commedia*.

2. Modified versions of this system have been employed by the Canterbury Tales Project and by Dorothy Severin and Fiona Maguire for their Electronic Corpus of 15th Century Castilian Cancionero Manuscripts (http://cancionerovirtual.liv.ac.uk/main-page.htm).

3. My involvement in the *Commedia* Project was made possible through STEMMA, a project funded by the Leverhulme Trust between 2000 and 2003.

4. This DVD-ROM includes a detailed article by Prue Shaw, "General Transcription Note," which describes the transcription system in detail. For general information about the transcription system employed by the *Commedia* Project readers are referred to that article. Here I choose illustrative examples which show how these guidelines developed.

5. At the end of this article, we have included a list of the elements used in the edition. The use of this system with other texts is the subject of a forthcoming article.

6. Our encoding of this passage can be found below.

7. For example, Codex Sinaiticus shows changes and corrections made by many correctors over several centuries.

8. All references to XML-TEI are to P5 (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html).

9. The <add> and <del> elements are the standard TEI recommendation for added and deleted text, as described in P5. See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COEDADD, under the heading 3.4.3 Additions, Deletions, and Omissions.

10. Peter Robinson and I reached this decision after several conversations. Robinson's involvement in the original TEI guidelines and his responsibility in the original implementation of <add> and <del> were invaluable in this new examination of their use.

11. A possible exception could be found in the use of the word deletion when used as a synonym of erasure. However, it is often the case, and particularly within a context of manuscript culture, that the cancellation of a text can be expressed in a variety of forms.

12. Notice how all these processes are mediated by an agent (the editor/transcriber/ reader) who attributes a conventional meaning to them. Indeed, faced with the same set of circumstances different editors are very likely to interpret the same text as a deletion.

13. Here, the word original does not mean archetypal, instead it denotes the oldest word present in a

particular document.

14. This could also be <rdg type="c2">, <rdg type="c3">, <rdg type="c4">, etc., depending on the number of distinct scribes or correctors in a particular witness.

15. The word “del” could be understood as an alternative reading, rather than a correction.

16. In practice, the encoding system employed in this publication does not follow all the recommendations included in the internal guidelines. This is particularly noticeable in the treatment of the modification of minims. In the previous example, when the scribe modifies the word “sicuro” to read “maturo”, it is not strictly correct to say that the letter “m” was written over “si”, which is what the encoding seems to express. My recommendation for instances that involve the modification of minims was to make use of an entity (&i;) within the element <rdg type="lit"> to separate them. The previous example would have been expressed as:
<app>
<rdg type="orig">sicuro</rdg>
<rdg type="c1">maturo</rdg>
<rdg type="lit"><s type="rp"><s type="cow">s</s>&i;&i;</s>&i;<s type="il">a</s><s type="rp"><s type="cow">c</s>t</s>uro</rdg>
</app>

17. In this instance, the encoding system suggested to describe it is as follows:
<app>
<rdg type="orig">prima</rdg>
<rdg type="c1">priua</rdg>
<rdg type="lit">pri<hi rend="er"><hi rend="ud">&i;</hi></hi>&i;&i; a</rdg>
</app>
This describes the state of the document in which one of the minims of the “m” was both underdotted and erased thus producing the new reading. Such change would have represented an insurmountable difficulty for our previous encoding system, but we can now encode the change by using minims within the <rdg type="lit">. This might not seem like a very big leap, but it implies a different kind of thought, a different conception of the final purpose of these transcriptions and their encoding.

18. While the Bergen edition of the Wittgenstein *Nachlass* edition distinguishes the variant states of the text as does our encoding, it does not provide a representation of the text of the document, as we do. At the time of writing, a TEI workgroup on encoding of genetic manuscript transcription is considering the matter. While their work is not yet complete, a preliminary report at http://users.ox.ac.uk/~lou/wip/geneticTEI.doc.html#index.xml-body.1_div.1_div.1 shows that this group is addressing the same distinction between “the text of the document” and the “variant states

of the text". However, this distinction is differently expressed, as between simply "document" and "text" (or, in German, between "befund" [record] and "deutung" [meaning]). As explained below, this is more than a difference of expression. Further, the system they offer suggests an entire separation of the transcription of the two levels. Thus, one would transcribe the "document" into one structure; the "text" into another, with complex links between the two. This is rather more complicated than our scheme, which focuses only on places of variation within a continuously-written document and seeks to include all encodings within a single encoding of that document. The range of situations addressed by the workgroup is far wider than encountered in manuscripts of the *Commedia*. However, it can be argued the solution here proposed, where the variant states of the one text present in one document in a single structure are encoded, has considerable advantages.
It is significant that the first example given in the workgroup document, and which is used to illustrate the complete separation of transcripts of "document" and "text", is of a diary entry which contains two separate texts: one beginning "Feed birds in the park today..", and a second, written at right angles to the first, beginning "Samaria is a Greek brand of water..". In the terms we use, these are not variant states of the text at all: they are actually quite distinct texts, which happen to be written on the one piece of paper. Here the difference between the distinction this paper offers, between the text of the document and the text or texts which might be constructed out of it, and between simple "document" and "text" as offered by the workgroup, becomes important. For our work, our distinction serves us well. [Note contributed by Peter Robinson]